Chip-Scale Energy and Power… and Heat

Prof. Paul Hasler

Electrical and Computer Engineering Department,
Georgia Tech University

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Maximizing Computational Capability with Minimal Power	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Georgia Institute of Technology,Electrical and Computer Engineering Department,Atlanta,GA,30332**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
MTO (DARPA Microsystems Technology Office) Symposium, 2009 Mar 2-5 2009, San Jose, CA. U.S. Government or Federal Rights License.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**22**	

Maximizing Computational Capability with Minimal Power

Paul Hasler

Professor, Georgia Institute of Technology

DARPA activity:
ISP, CT2WS, SyNAPSE, Healics, TEAM

Power Efficient Computing

Portable Devices
- battery powered
  (or less)
- larger systems
  minimize battery size / weight

Get as much computation
as possible…

Custom Analog ~ 1000 – 10000
more efficient than Custom Digital
(Mead 1990)

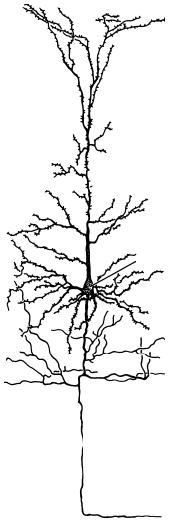
Analog (VMM): 10MMAC/ μW
Digital: 4 MMAC / mW (DSP)

Useful Analog must be
Programmable / Configurable

Cortical Neurons
- 1000's of inputs,
- 1000's of channel populations,
- one output

Equivalent computation ~
400MMAC / neuron
(no learning / growth)

~ roughly 20pW / neuron



400MMAC / neuron at 20pW…
digital is quite far away (100mW)
analog VMM closer (100μW)
analog HMM / dendrites get close…

~ 200TMAC
< 500 neurons
~ 40kW (comp) with 2000 DSPs

Modern System Design

Fixed function Digital

(1837)

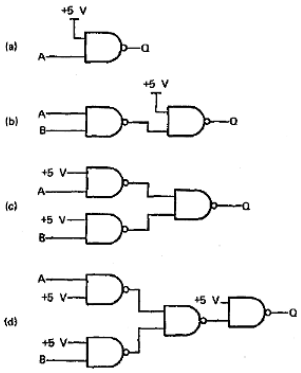
↓

Fixed function Analog

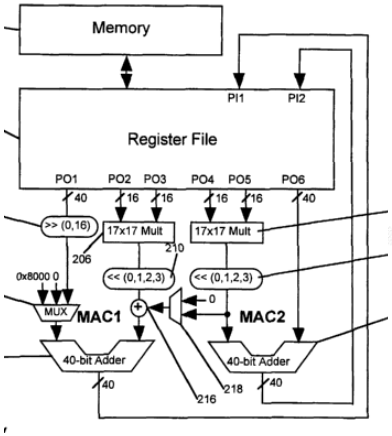
↓

Programmable Digital (Mixed mode)

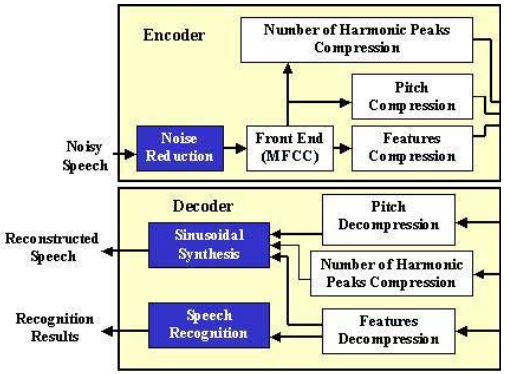
Design at gate level



Design at Multipliers and Adders



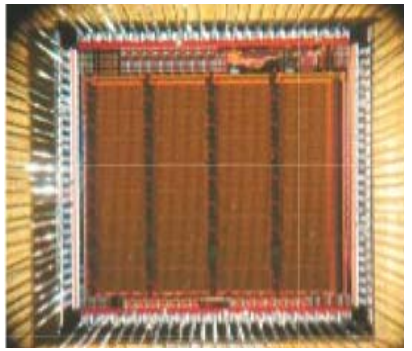
Design at Basic Algorithms



Vector-Matrix Multiplication
Frequency Decomposition
Adaptive Filters
Classifiers (NN, GMM, HMM)

When building analog systems,
we expect to build primitives at the basic algorithm level....

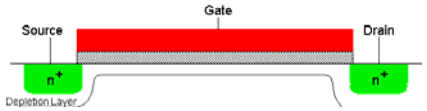
Analog = programmable and configurable.



How to get enough analog engineers

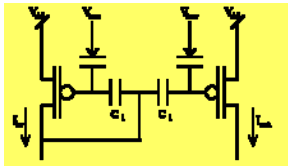
Hierarchy is a key ingredient to the success of the digital circuit, and, until recently, one reason why large analog designs have been difficult

Levels of Energy Efficiency



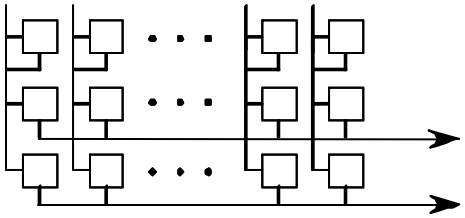
Subthreshold Transistor Operation

Highest throughput / amount of power



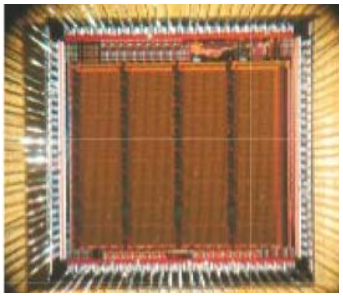
Programmable Circuits (FG transistors)

- Eliminate mismatch
- Programmability



Analog Signal Processing

- ~ x1000 improvement in power efficiency



Configurable Signal Processing

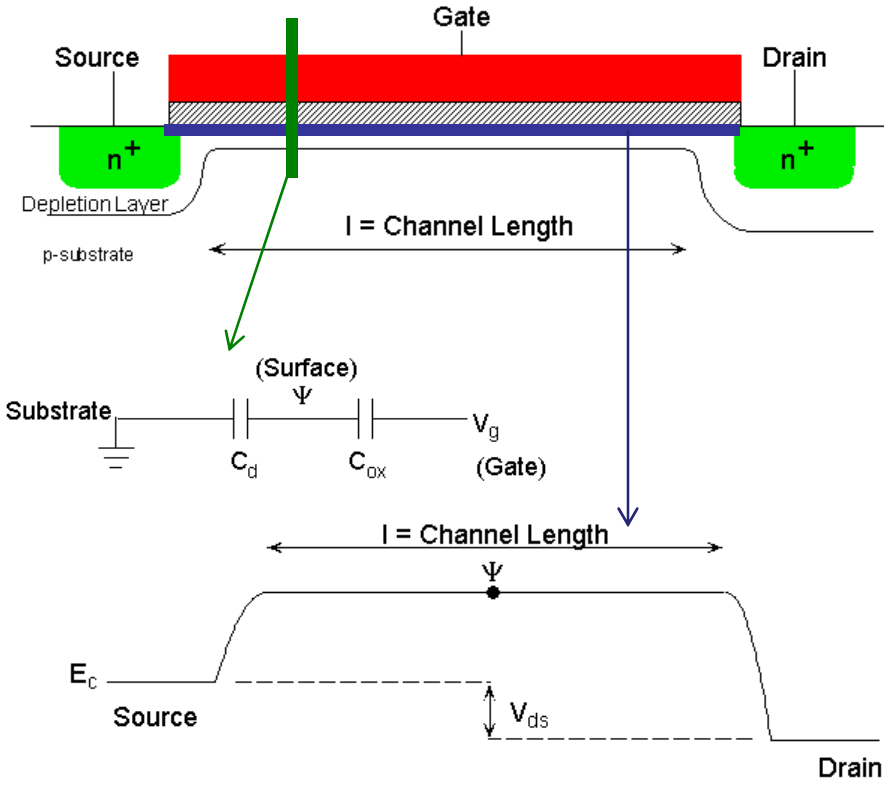
- Wide accessibility

Moving analog approaches /conceptual framework to a system design approach, similar to digital's system transformation in the 1970's / 80's.

- Large need for tools to compile / program these systems.
- Link most "useful" at system /sig processing level
- Education / training / foundational theory is critical for designing.

These techniques open further opportunities to utilize / explore biologically inspired techniques

MOS Transistor Derivation

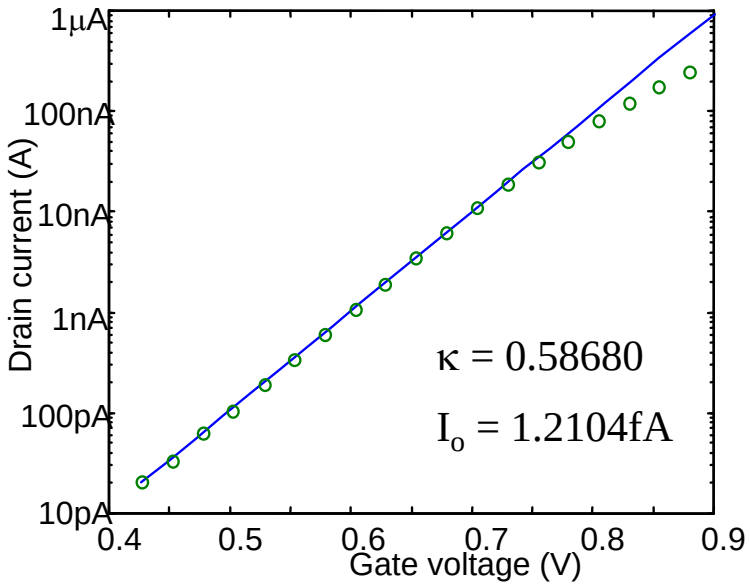


$$I_{ds} = I_o\, e^{\kappa V_g / U_T}\left(e^{-V_s / U_T} - e^{-V_d / U_T}\right)$$

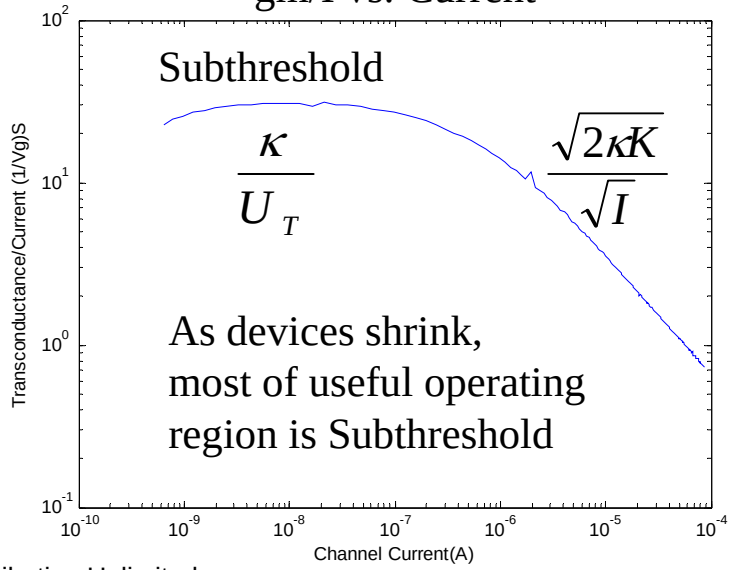
$$= I_o\, e^{(\kappa V_g - V_s)/ U_T} \qquad (V_{ds} > 4\, U_T)$$

"Saturation'

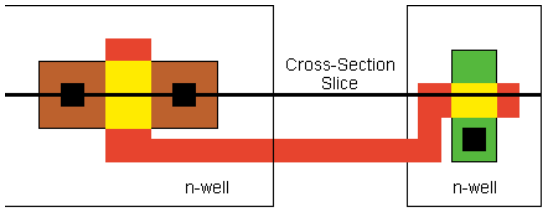
Mismatch is significant: 10mV V_T shift
~ 50% bias current variation

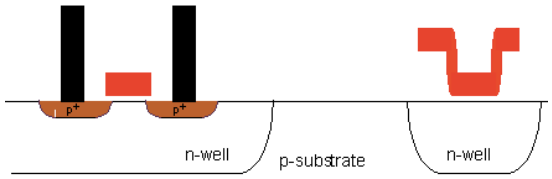


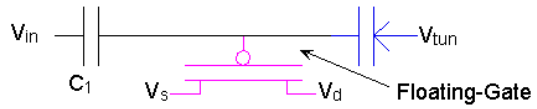
gm/I vs. Current



Programmable Analog Transistors



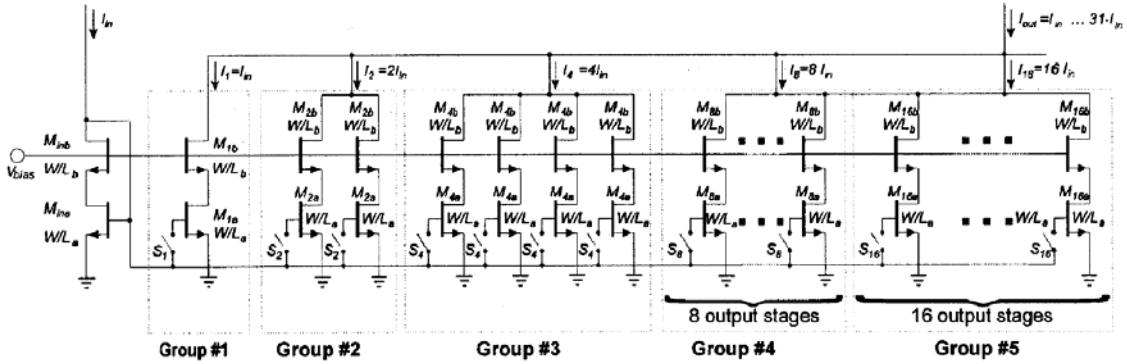


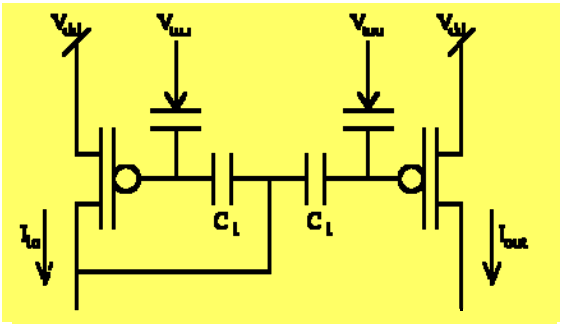


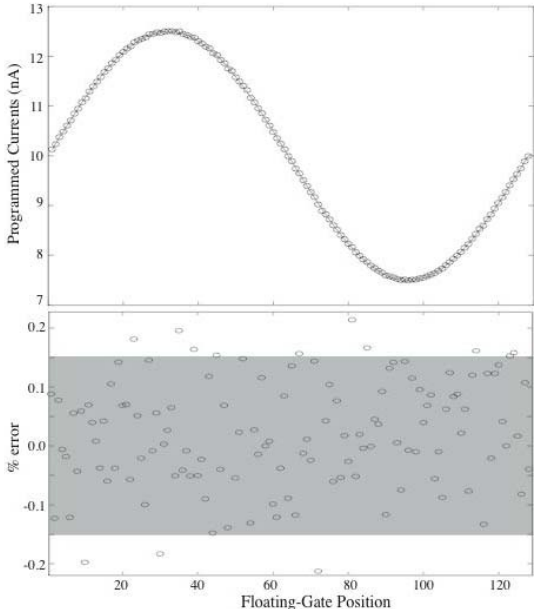
- Standard CMOS
- Data retention:
  < 5μV (0.5μm) (10 year, 300K)
- Apps: Filters, Data converters, Regulators, etc.

Accuracy ~ 0.1% between
100pA - 1μA ~ $10e^{-}$

Write degradation (100μC):
V_{tun} increase less than 25%
V_{inj} negligable change
(100μC is $>10^9$ complete FG rewrite)

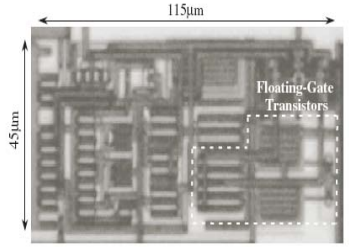


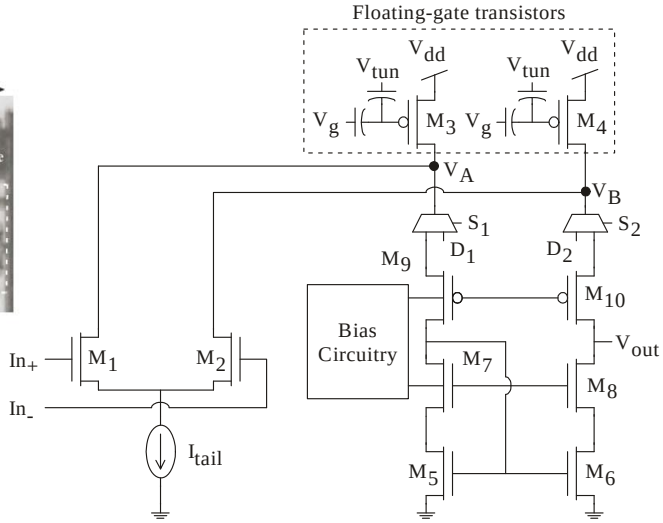


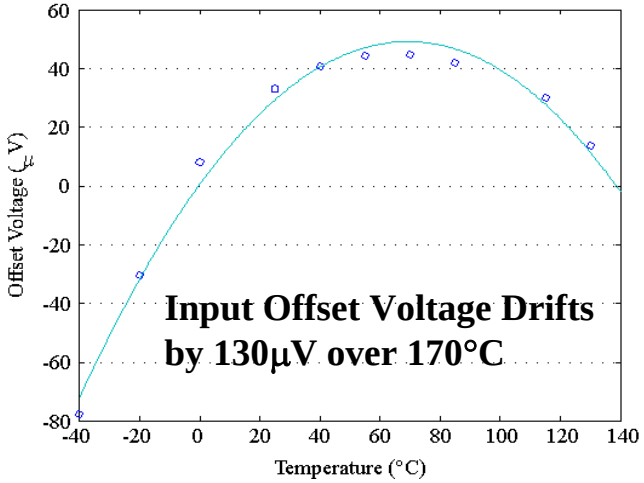


Otherwise, need a DAC at every parameter and/or memory, etc.

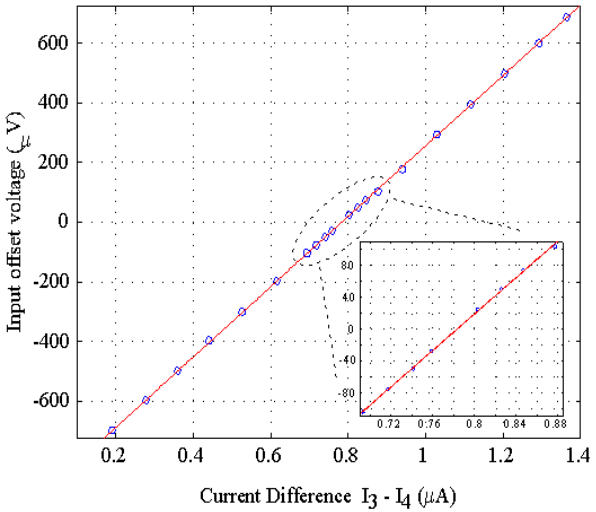
Industrial Quality Programmable Analog ICs







Input Offset Voltage Reduced to ±25µV

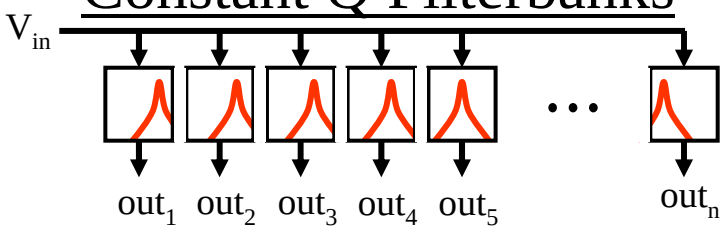


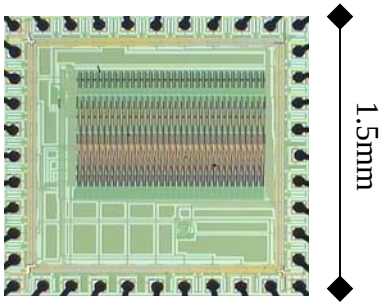
Parameter	Value
Supply Voltage	$3.3V$
Technology	$0.5\mu m$ CMOS
Input Common Mode Range	$1.2V - 3.1V$
Output Voltage Swing	$0.2V - 3.1V$
Input Offset Voltage	$\pm 25\mu V$
Offset Voltage Drift with Temperature	$130\mu V/170°C$
Offset Voltage Drift @ $25°C$ for 10 yrs	$0.5\mu V$
Open Loop Gain	$63dB$
Unity Gain Bandwidth @ $C_L = 20pF$	$10MHz$
Phase Margin	$60°$
Common Mode Rejection Ratio	$73dB$ (Simulation)
Power Supply Rejection Ratio	$77dB$ (Simulation)
Input Referred Noise (rms)	$8.9\mu V$ (Simulation)
Slew Rate	$5V/\mu s$
Settling Time (10 Bit) for $100mV$ Step	$105ns$
Power Dissipation (Incl. Buffer)	$8.25mW$
Area (Excl. Buffer)	$115\mu m \times 45\mu m$

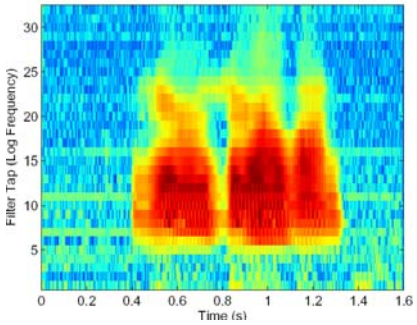
V. Srinivasan, G. Serrano, J. Gray, and P. Hasler, CICC 2005, pp. 739-742. (Best paper CICC 2005)

Analog Signal Processing Techniques

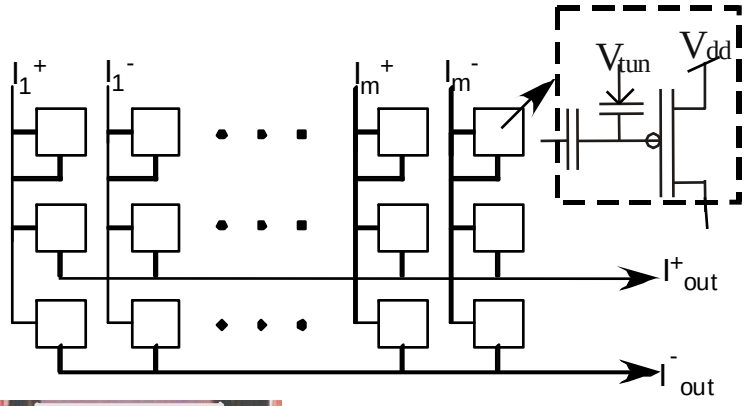
Constant Q Filterbanks





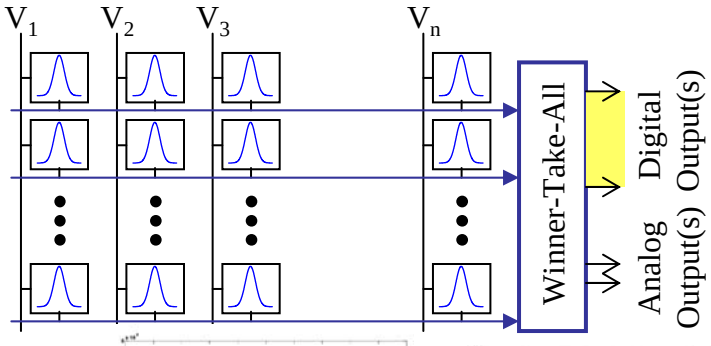


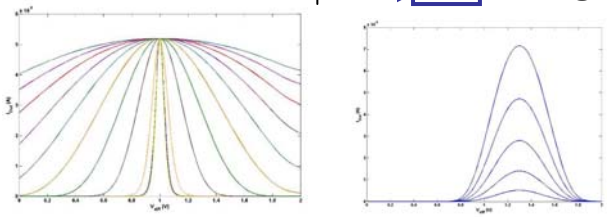
Vector-Matrix Multiplication



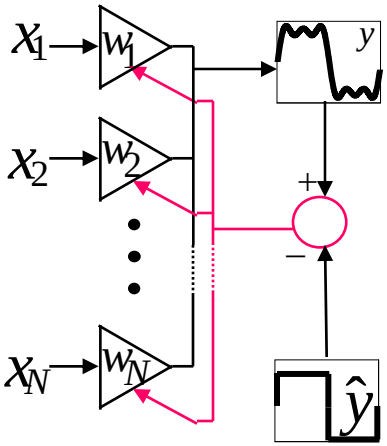
128x32 VMM (0.5µm)
< 1mm^2

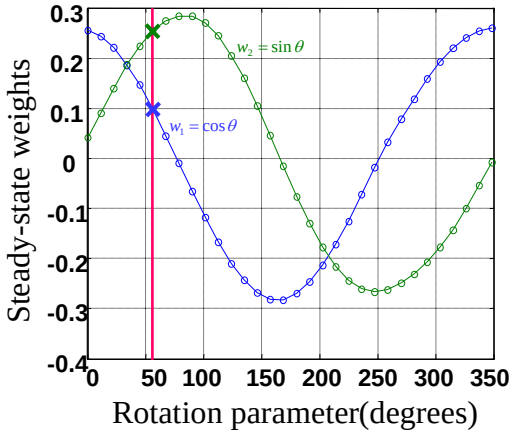
Gaussian Mixture Models / VQ



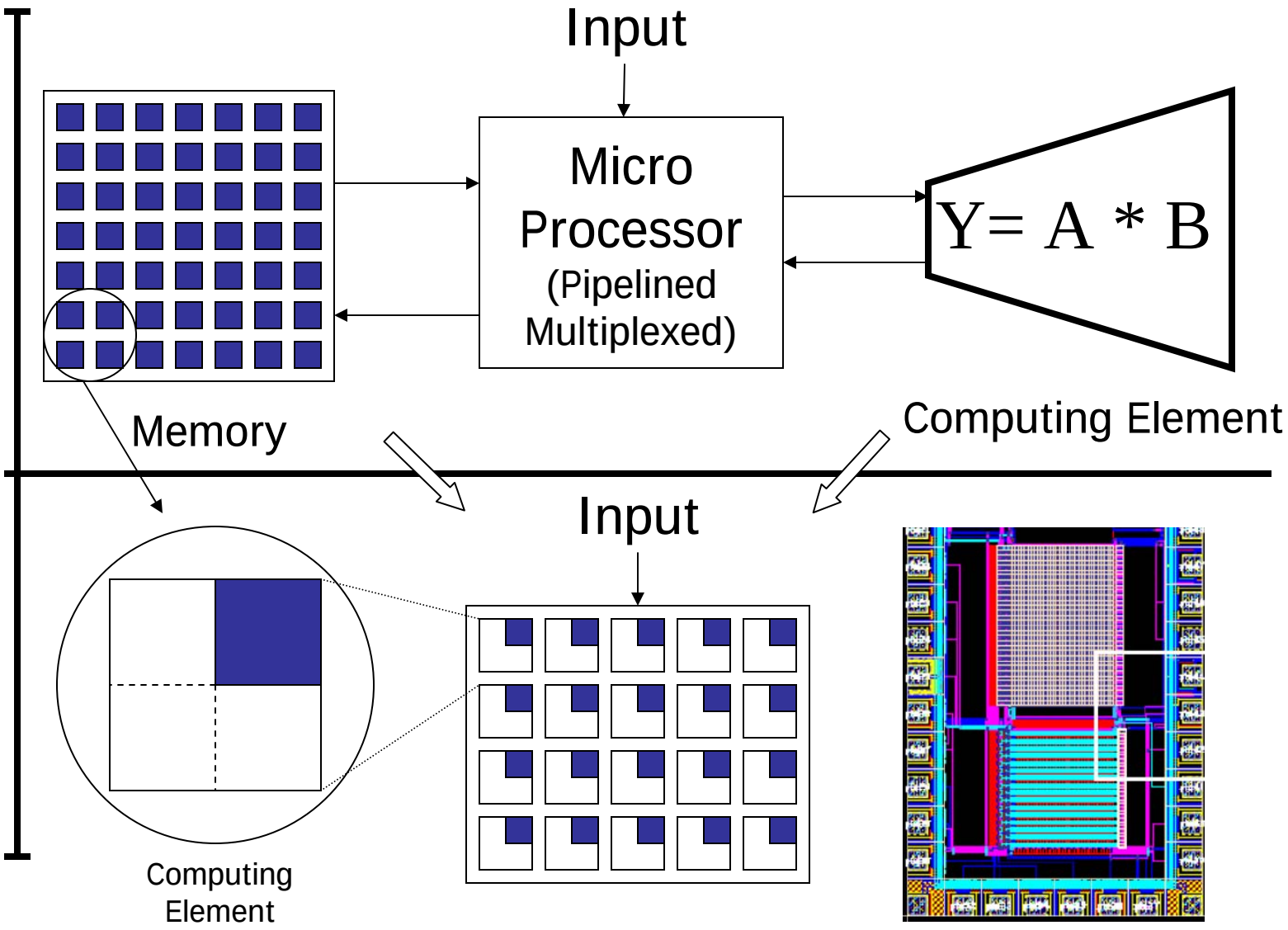


Adaptive Filters

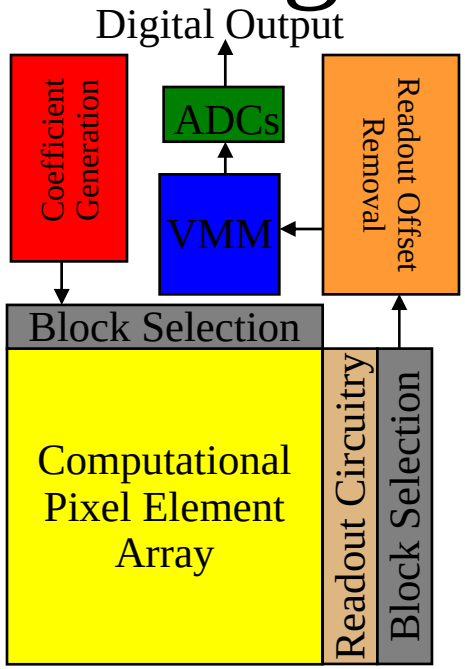




Computing in Memory



Programmable Transform Imager



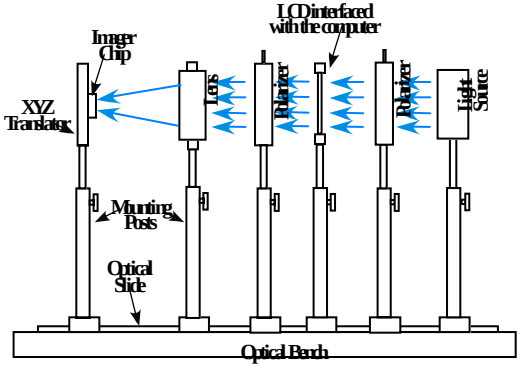
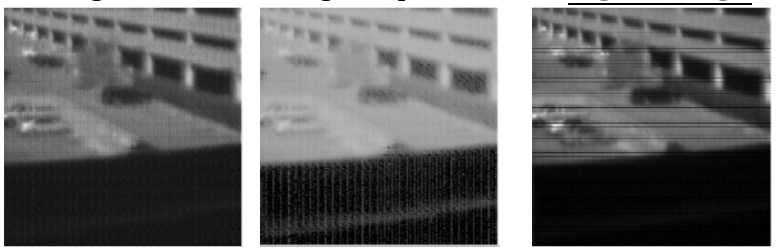


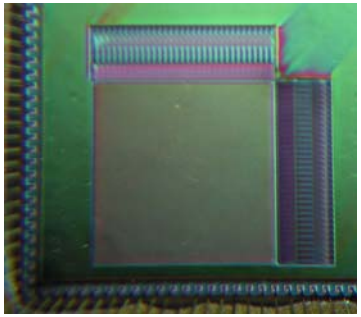
Image Read | Log compress | DCT -> IDCT

Image Read | Log compress | DCT -> IDCT

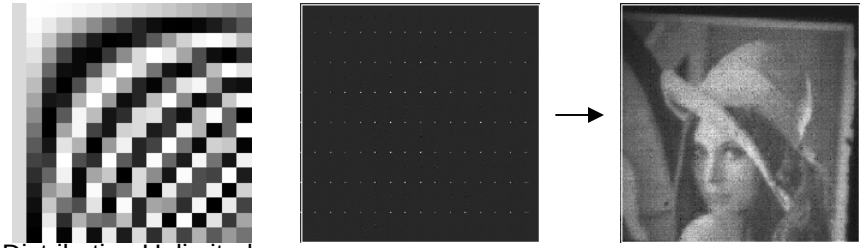


Technology	0.35µm CMOS
Array size	256 x 256
Pixel size	6 µm x 6 µm
Fill factor	38%

Original | Measured | Edge Enhan

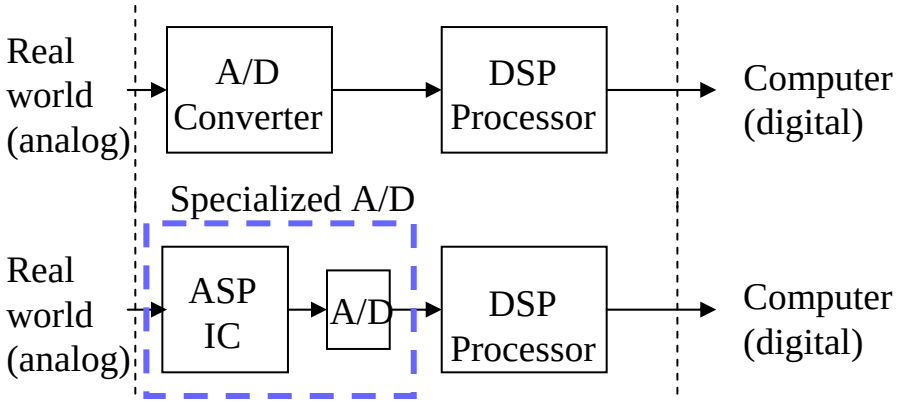


DCT Matrix | 2D DCT | Reconstruct



Die area =
4.5mm x
4.5mm

Analog--Digital Signal Processing

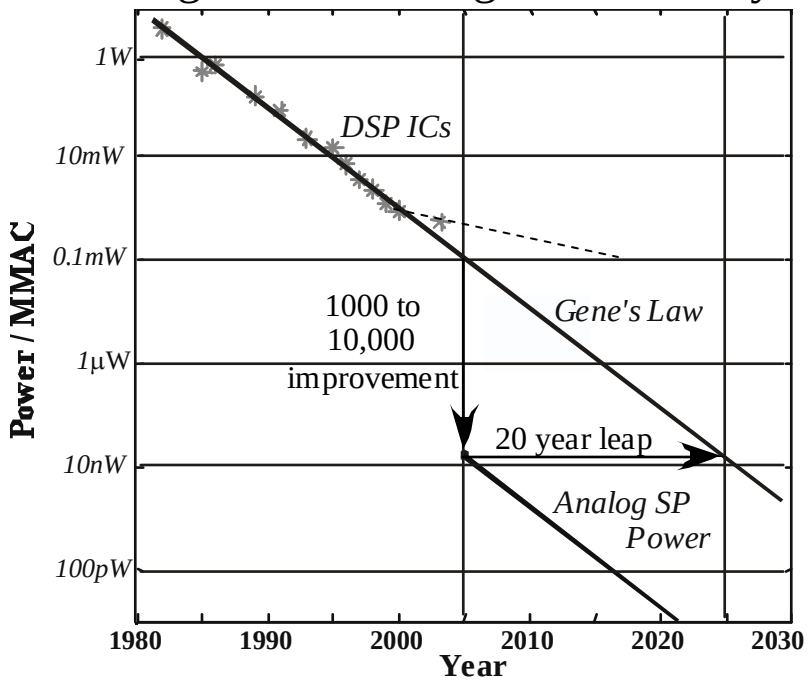


CADSP = Cooperative Analog—Digital Signal Processing

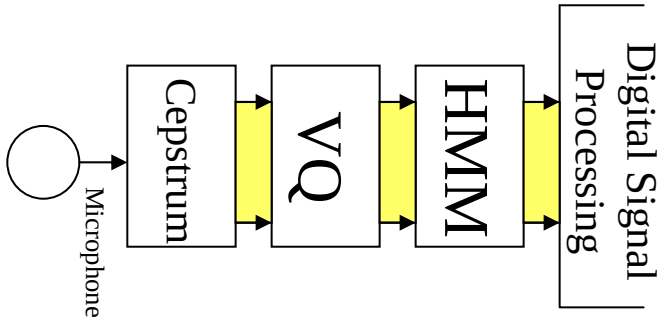
Custom Analog ~ 1000 - 10000 more efficient than Custom Digital (Mead 1990)

- Analog (VMM): 10MMAC/ μW (= 10TMAC / W)
- Digital: 4 MMAC / mW (DSP)

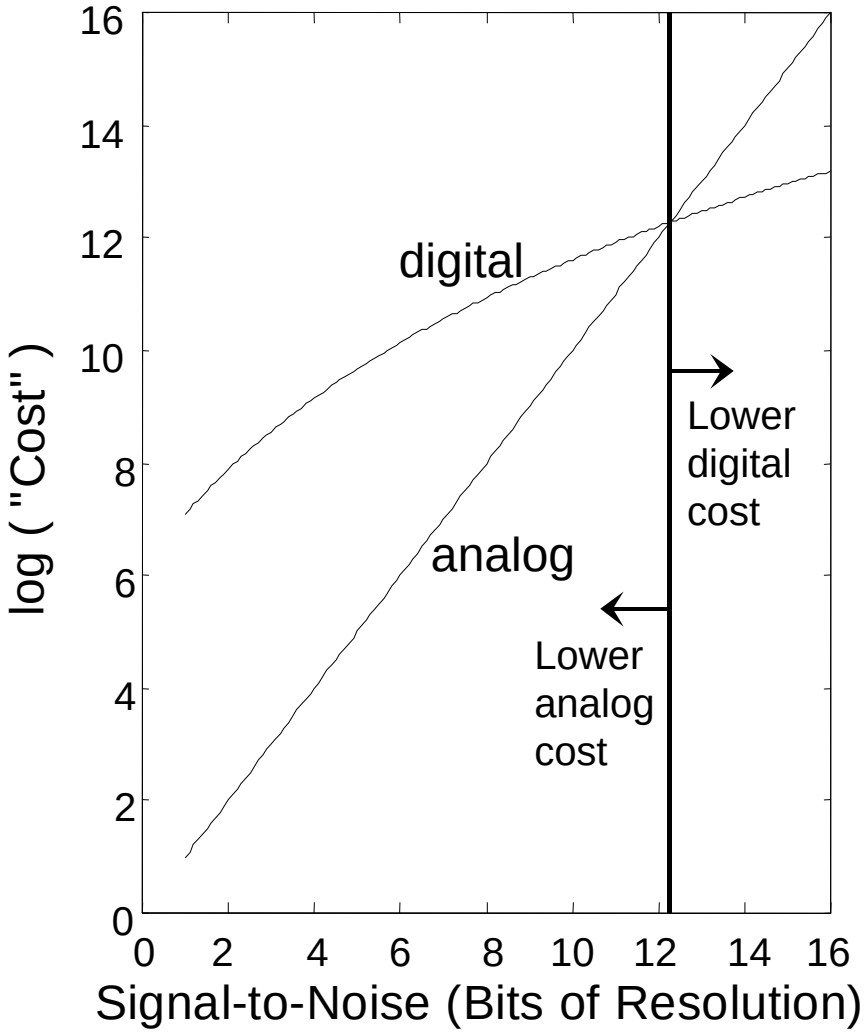
Digital and Analog SP Efficiency



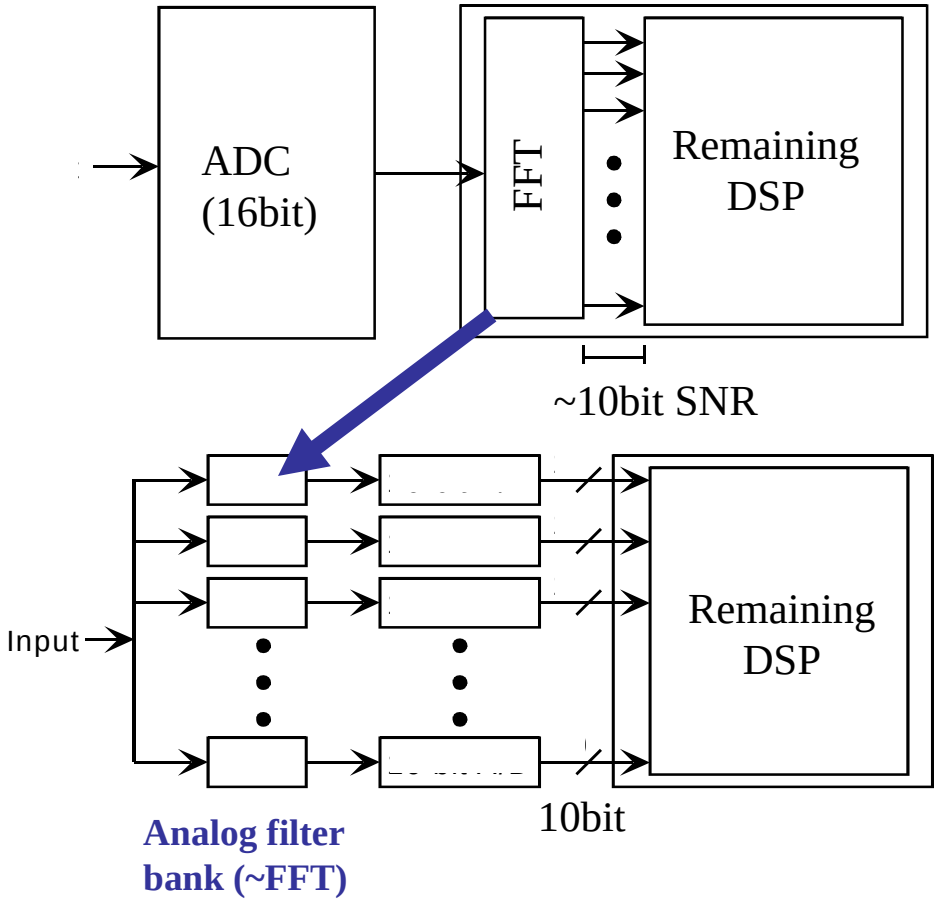
Computation	MMAC/μW	Ratio to digital
LowPowerDSPs	0.02 to 0.002	1
Analog VMM	1 to 30	1000
Analog Filterbanks	30 to 1000	10000
Analog VQ	1 to 10	300
Analog HMM	>1000	> 100000



Resolution for Analog / Digital Tradeoffs

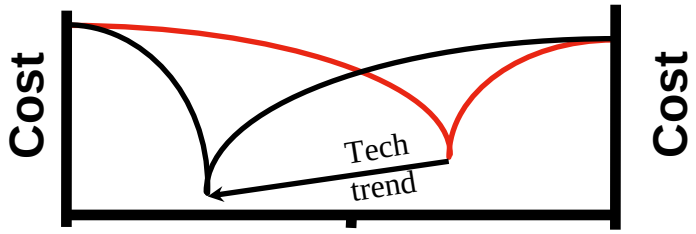


[Vittoz95, Sarpeskar98]



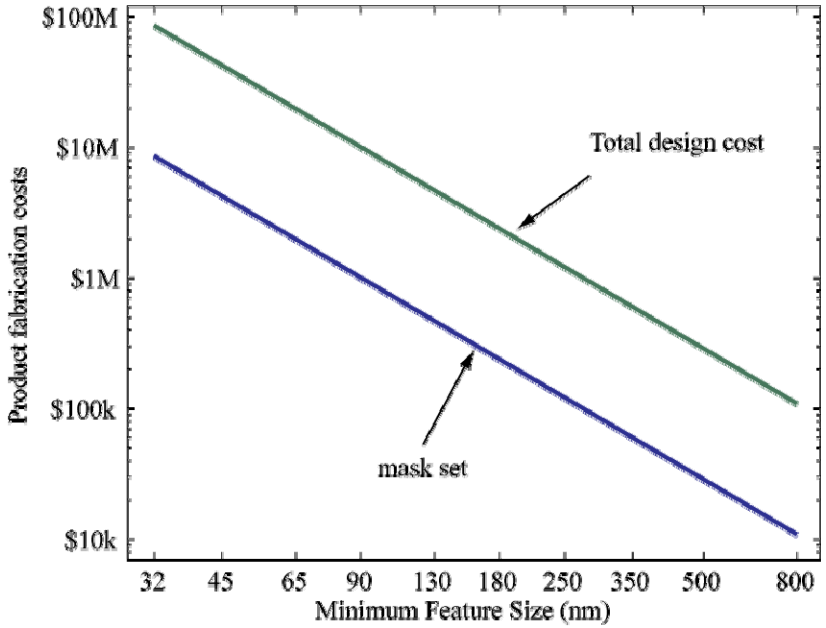
[Kucic, et. al. 2001]

Reconfigurable Signal Processing



100% S/W ⟷ **100% H/W**

(Programmable) (Fixed Function)



Innovation and Process Scaling moves solutions towards programmability and reconfigurability

FPGAs – Large Configurability

Power: Just MAC engine
around 2-10MMAC/mW
Baseline static power ~ 0.5W to 1 W
Signal routing power / memory: ?

DSPs – Low Power Processing
- cell phones
(processing < 30mW average)
- hearing aids (1 mW levels)
(AMI / DSP factory)

Power: 54C series – 4MMAC/mW

Power does not include comm off chip
(i.e. accessing memory)

Power = ½ C Vdd^2 f for CMOS
Chip to Chip (10pF load min, 2.5V):
32uW/Mbit (dynamic)

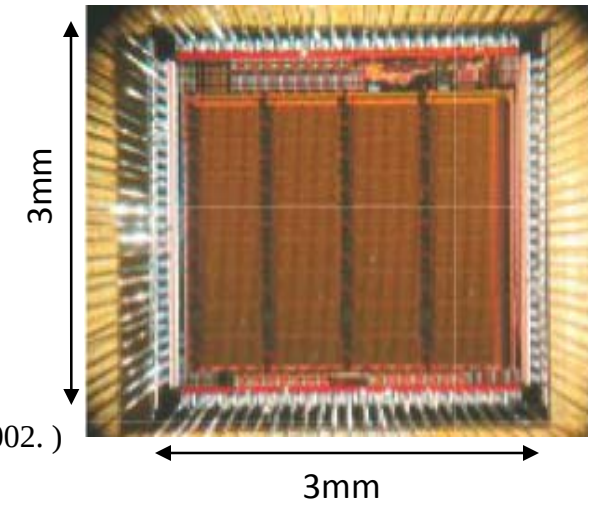
Obtaining data for 4MMAC computation ~ 4mW

Moving towards Configurable Analog

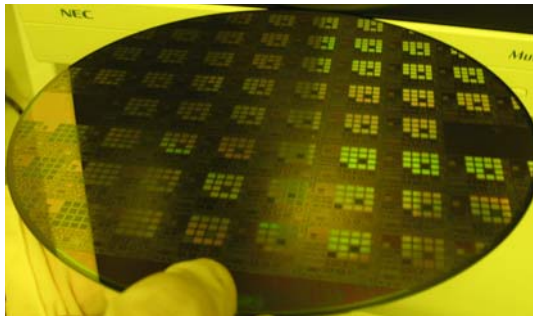
Useful Analog must be Programmable / Configurable

FPAA =
Field Programmable
Analog Arrays

Can be a prototyping tool,
early devices, or
final application



Needing large number
(>1000) of FPAAs



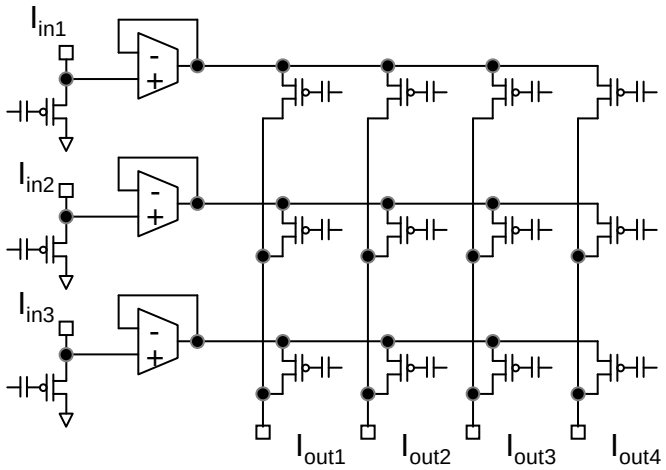
- RASP 1.x (2002)
  (T. Hall, P. Hasler, et. al, FPL, Sept. 2002.)

- RASP 2.x:
  RASP 2.5, 2.7: 2004-2007
  (C. Twigg & P. Hasler, CICC, 2006)
  - >50,000 Prog. Analog Devices
  - Used by > 100 Eng

  RASP 2.8x: 2008-
  (A. Basu, et. al, CICC, 2008)
  - Used by > 50 Eng and growing

  RASP 2.9x: 2009-

Switches are not dead weight



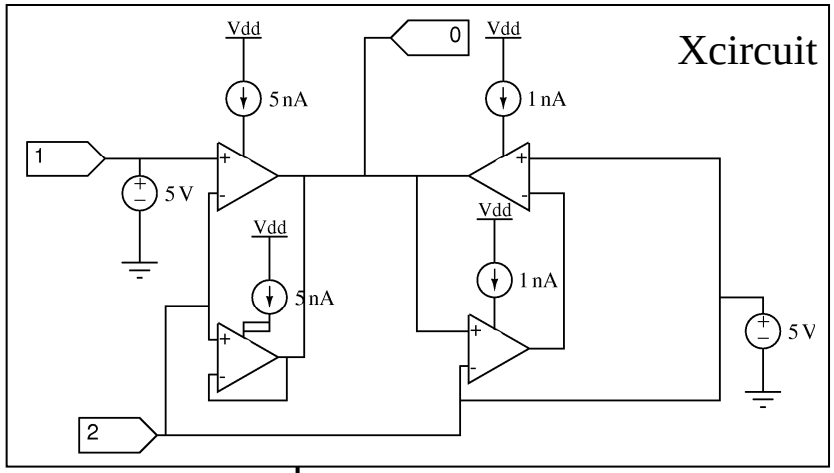
Custom versus FPGAs: x2-3 speed, x10 area, x100 power
Custom versus FPAAs: < x2 speed, < x2 area, < x2 power

Next Questions on FPAAs

FAQ on Large-Scale FPAAs

- Design time similar time for FPAA targeted and custom ICs
- Size can be similar to custom (programmable caps / I)
- Noise levels are similar to custom design
- Similar speed as custom upto routing fabric speed (~10-20MHz in 0.35um CMOS)
- Power levels often similar to custom solutions
- Techniques scale (~ ideal CMOS rules) with process shrink

Node (nm)	Prog #s (M)	TMACs
350	4.0	1
90	64.0	64
45	256.0	512



Extract

Spice Netlist → Simulate / Verify

Compiler (RASPER)

Computer (MATLAB) ↔ USB

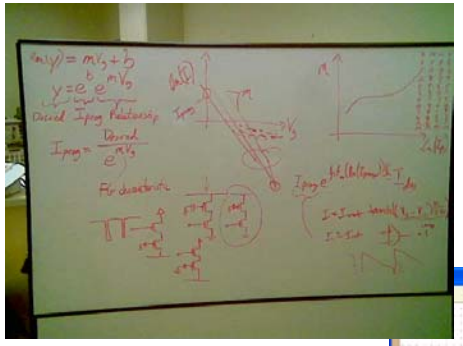
- Neuromorfix to commercialize FPAA technology

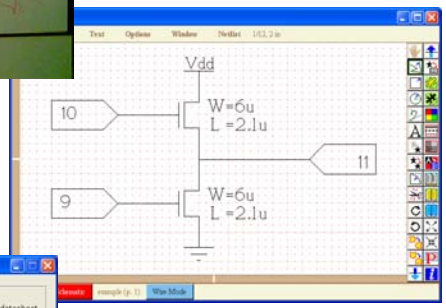
Compiled circuits include:

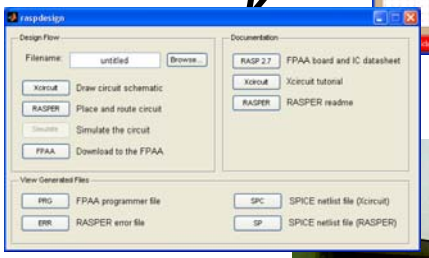
n-th order filters / filterbanks, Capacitive summation / differencing, Ramp ADC, Algorithmic and Sigma-Delta ADCs, MP3 encoder, WTA, Analog Distributed Arithmatic, HMM classifiers, Van-der-pol Oscillator

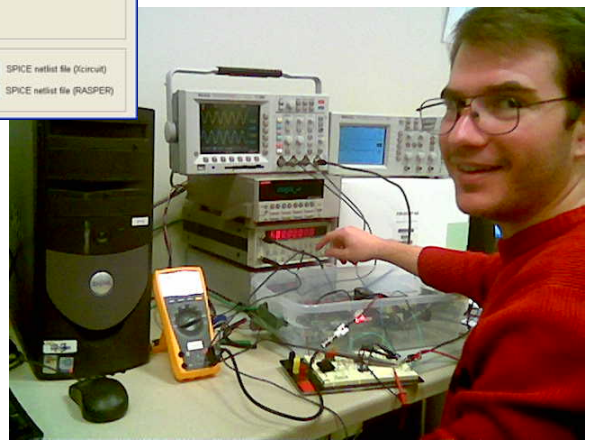
Rapid Prototyping using FPAAs

RASP 2.7 PhotoReceptor Response

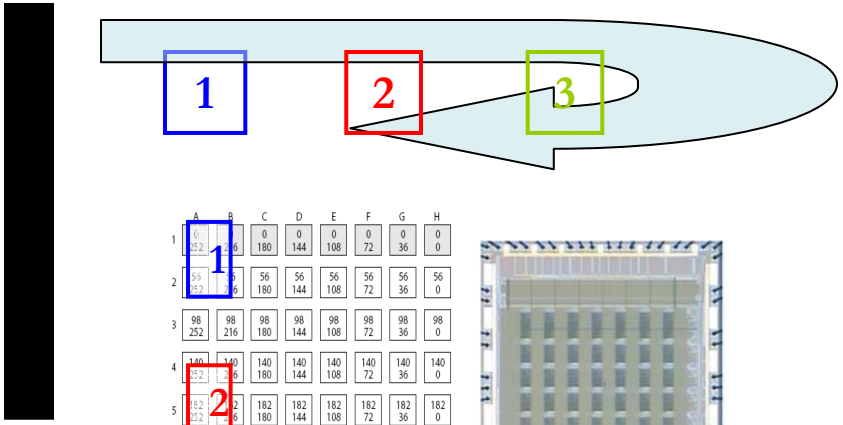


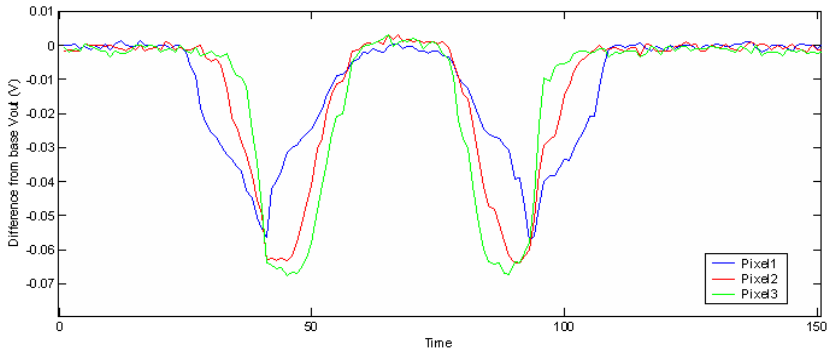






Paper Strip





FPAA Workshops (RASP 2.8x)

LA Workshop
USC Campus, May 2-7, 2008

>30 Participants.

CO Workshop
Telluride, July 2008

> 20 Participants.

ATL Workshop, Oct 2008

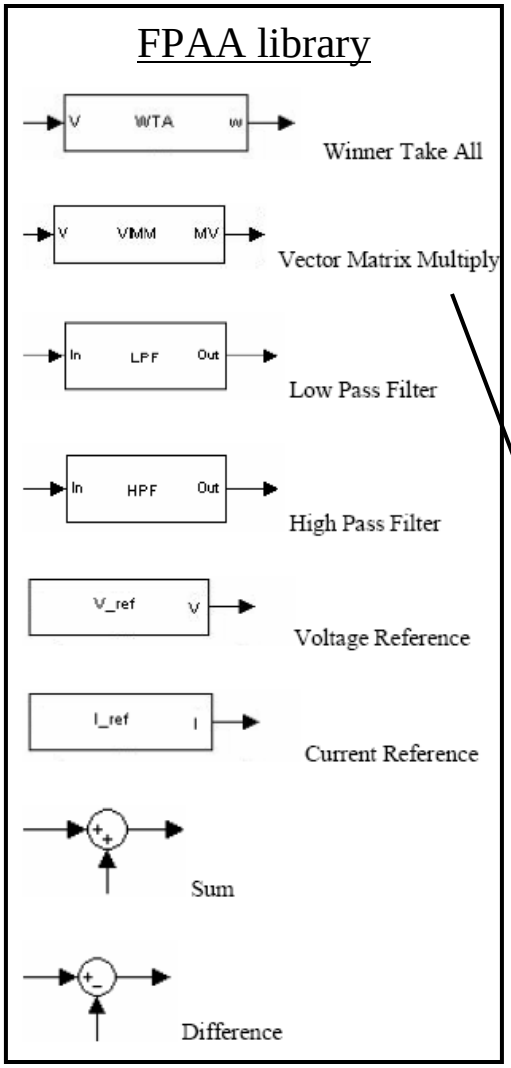
>25 Participants.

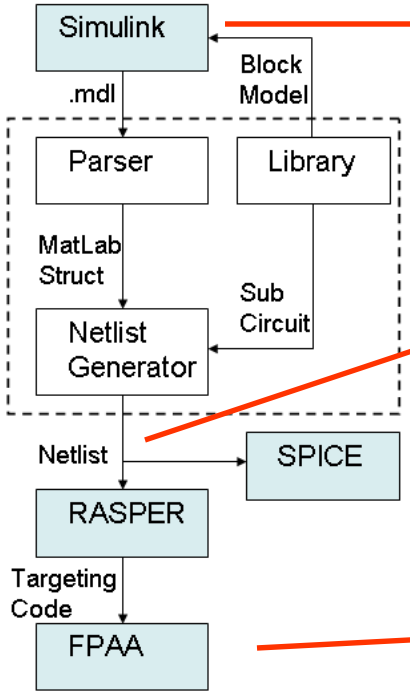
Other workshops being planned:
Boston, SF, Orlando, DC?

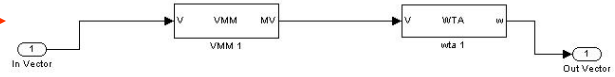
GT Neurmorophic Classs
(Fall2008, >20 students)

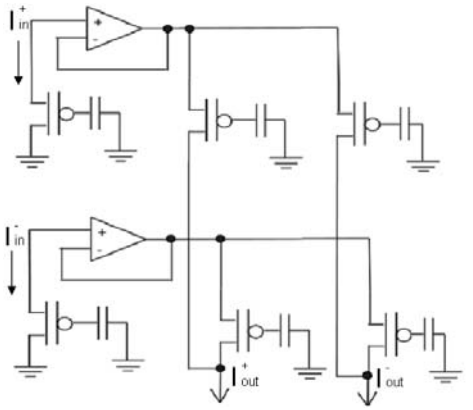
Education / training / foundational theory is critical for designing.

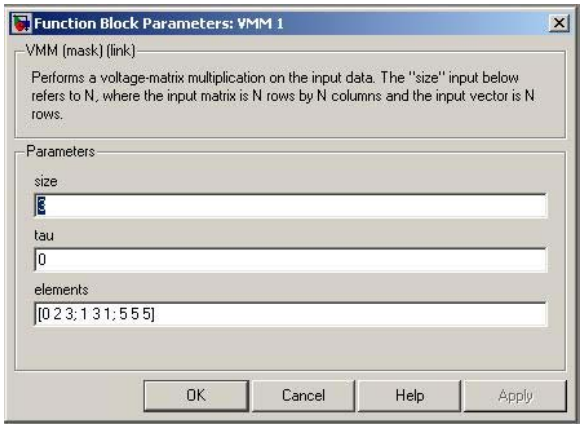
Simulink FPAA Tool

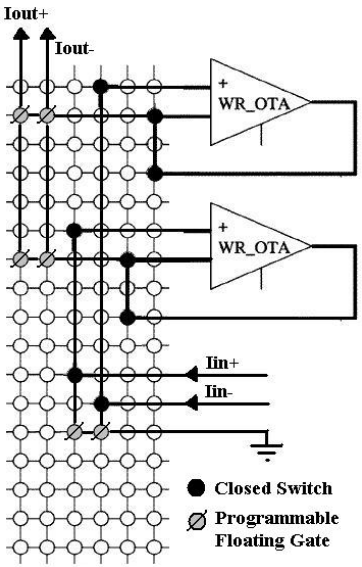








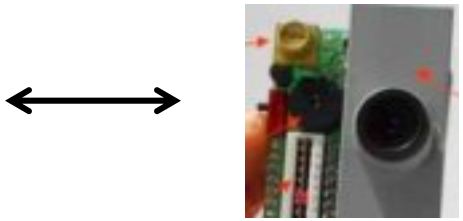




Petre, et. al, ISCAS 2008]

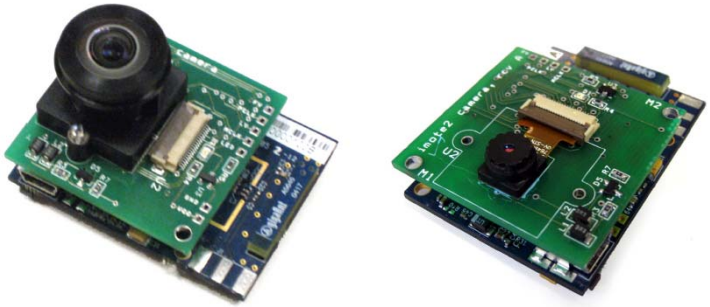
Getting higher power efficiency: Neuromorphic Engineering

400MMAC / neuron at 20pW
vs. digital (100mW)
and analog SP (100μW)



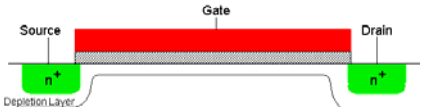
- Neuromorphic processing = event-based processing
  uses power only when useful signals are present
  ("always on" in sensors or further processing)

Programmability and Configurability empowers neuromorphic design towards useful applications in a reasonable timeframe.
- Address Event Representation (AER) / FPGAs
- FPAAs / FG devices –
  ~ sizes of largest custom neuro ICs



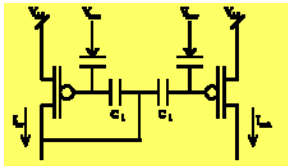
Can model pryamidal cells in configurable fabric in ~1mm^2 area with realistic channel, dendrite, and synapse elements (power in nW level and decreasing)

Levels of Energy Efficiency



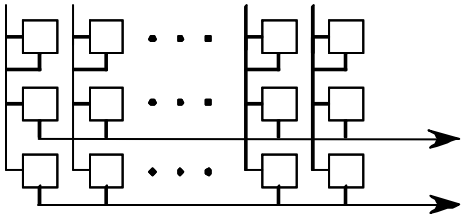
Subthreshold Transistor Operation

Highest throughput / amount of power



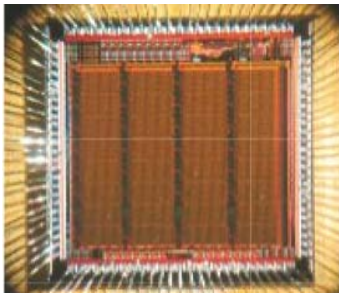
Programmable Circuits (FG transistors)

- Eliminate mismatch
- Programmability



Analog Signal Processing

- ~ x1000 improvement in power efficiency



Configurable Signal Processing

- Wide accessibility

Moving analog approaches /conceptual framework to a system design approach, similar to digital's system transformation in the 1970's / 80's.

- Large need for tools to compile / program these systems.
- Link most "useful" at system /sig processing level
- Education / training / foundational theory is critical for designing.

These techniques open further opportunities to utilize / explore biologically inspired techniques

DARPA activity:
ISP, CT2WS, SyNAPSE, Healics, TEAM